

Towards Flexible, Natural, Efficient Interaction for Conversational Talking Face Generation

Baiqin Wang^{1,2}, Sen Chen^{1,2}, Jiankuo Zhao^{1,2}, Xiangyu Liu^{1,2},
Zhen Lei^{1,2,3,4}, and Xiangyu Zhu^{1,2*}

¹ MAIS, Institute of Automation, Chinese Academy of Sciences

² School of Artificial Intelligence, University of Chinese Academy of Sciences

³ CAIR, HKISI, Chinese Academy of Sciences ⁴ SCSE, FIE, M.U.S.T
{wangbaiqin2024, zhen.lei, xiangyu.zhu}@ia.ac.cn

Abstract. Conversational talking face generation has recently attracted increasing attention, aiming to synthesize interactive talking videos where characters speak, listen, and respond dynamically to each other. This task presents three core challenges: 1) Flexibility: enabling multi-round dialogues with an arbitrary number of participants; 2) Naturalness: maintaining coherent motion and appropriate non-verbal feedback throughout the interaction; and 3) Efficiency: achieving real-time generation and low computation overhead for long-term continuous online conversation. Despite recent advances, existing methods still fall short in balancing all three requirements. To bridge this gap, we introduce InterTalk, a novel and efficient framework designed for highly interactive conversational talking face generation. Built upon a motion-based architecture, InterTalk supports real-time conversation synthesis. Our method achieves strong flexibility by explicitly modeling multi-round conversational dynamics among each participant, eliminating constraints on their numbers. To enhance interactivity, we incorporate motion feedback from multiple participants and introduce an iterative generation strategy for more natural behaviors. Besides, we disentangle motion into several facial components, enabling targeted refinements for natural response such as precise lip-sync and realistic eye-blinking. Finally, we construct a new multi-person conversational dataset and enrich it with 3D face-based data augmentation. Extensive experiments demonstrate that InterTalk achieves superior interaction quality while maintaining real-time performance at 30 FPS.

Keywords: Talking Face · Interaction · Efficiency · Conversation

1 Introduction

Talking face generation aims to create expressive facial animations with lip movements synchronized to speech. While existing methods [3, 10, 12, 30, 32] mainly

* Corresponding author.

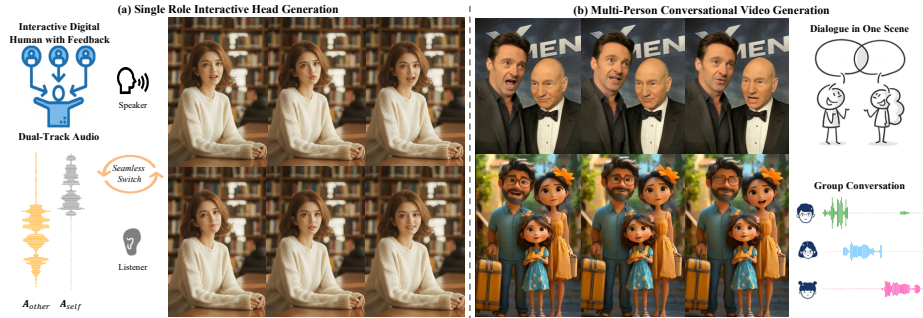


Fig. 1: Illustration of InterTalk. InterTalk generates conversational talking face with flexible and natural interactions in real-time. (a) For a single participant, it generates an interactive talking face capable of multi-round dialogues with seamless role switching. (b) For multiple participants, it produces coherent talking-face videos within a shared scene, capturing realistic group interactions.

focus on only speaking scenarios, there remains great potential in interactive conversational talking faces between participants [11, 14, 35]. Unlike single-speaker generation, this task emphasizes fluid interactions among characters, including non-verbal feedback such as nods or subtle expression changes. Realistic conversational agents can significantly enhance user engagement in applications such as online education, remote meetings and emotionally digital companions.

In this paper, we identify three key requirements for conversational talking face generation. *Flexibility* demands the ability to model multi-round conversations between numerous participants. *Naturalness* calls for coherent interactions, including smooth articulation by speakers and timely non-verbal feedback from listeners, along with seamless transitions between characters. *Efficiency* demands real-time performance with low computational cost, which is critical for long-term continues applications like virtual companion. These aspects collectively form the foundation for advancing conversational talking face generation.

Current approaches for conversational talking face generation can be broadly divided into two categories. The first category explicitly models conversational dynamics to improve efficiency and interactivity but still faces several limitations. Some methods [26, 31] generate speaker and listener motions separately, requiring manual role switching which hinders multi-round conversations, while others, like INFP [35], rely on dual-track audio but focus on single-character scenes without feedback from others, resulting in isolated motions and weak responsiveness, confined to single-role scenarios. Additionally, their lip-sync accuracy remains restricted by limited conversational data. The second leverages video generation models [2, 20, 29] to synthesize multi-person conversation videos in an end-to-end manner, such as MultiTalk [14], which adopts Wan 2.1 [29] as its backbone. However, these methods are computationally intensive and rely on extremely large-scale training datasets, which is not suitable for application that requires high throughput generation including companion digital human and live-streaming. In summary, existing works struggle to simultaneously fulfill the three essential requirements of flexibility, naturalness, and efficiency.

To achieve these objectives, we propose InterTalk, a framework designed for both single-role and multi-person conversational scenarios as shown in Fig. 1. For the single-role interactive head generation, InterTalk provides coherent non-verbal feedback and seamlessly transitions between speaking and listening states. When it comes to multi-person scenarios, our method supports dialogue generation for an arbitrary number of participants within a shared scene for group conversation. Built on a paradigm that explicitly models interaction dynamics, InterTalk first incorporates a **Responsive Context Encoder** (RCE) to integrate all participants’ motion and audio cues into a unified interactive latent representation. This representation is then utilized by the **Interactive Motion Generator** (IMG) to produce *Flexible* multi-round conversational motions for each participant. To enhance interactivity, our framework incorporates motion feedback from all participants and adopts an **Iterative Generation Strategy** that progressively refines each participant’s motion, ensuring *Natural* and life-like multi-party interactions. Moreover, as different facial components exhibit distinct behavioral patterns—for instance, lip motion closely correlates with audio while eye movement is more constrained—we **Disentangle Facial Motion** into several components and model them separately. This design yields more natural behaviors and provides fine-grained control, where lip-sync accuracy is enhanced by a sync-enhancer module trained on single-speaker data, and an eye-blinking enhancer together with a head-pose initializer further improve visual realism. With an efficient rendering pipeline, InterTalk synthesizes multi-participant talking videos in *Real-Time* with low computation overhead.

Given the absence of publicly available interaction data involving multiple participants, we introduce a newly collected multi-person talking face dataset to bridge this gap. To further enhance the diversity and scale of the training data, we employ a **3D Face-based Data Augmentation** that converts 3D FLAME [16] coefficients into implicit keypoint, which provides additional supervisory signals on motions during model training.

In summary, our contributions are as follows:

- We propose **InterTalk**, a novel and efficient framework that generating multi-round conversational videos with an arbitrary number of participants, which simultaneously satisfies three key requirements of **Flexibility**, **Naturalness**, and **Efficiency** in conversational talking face generation.
- To enhance the interactivity, we incorporate motion feedback and introduce an **Iterative Generation Strategy** that enables each character’s motion to faithfully reflect feedback from others. We also **Disentangle Facial Motion** into distinct components, leveraging their distinct patterns to achieve more natural interactions and fine-grained controllability.
- We construct a new multi-person conversational dataset and further enhance training through a **3D Face-based Data Augmentation**, which converts 3D facial coefficients into implicit keypoints for effective supervision.
- Extensive experiments demonstrate that InterTalk achieves state-of-the-art performance in conversational talking face generation, delivering flexible, natural, and efficient interactions between participants.

2 Related Work

2.1 Single-person Talking Face Generation

Talking-face generation has made remarkable progress in recent years. Early methods such as Wav2Lip [22] adopt GAN-based architectures [8] to synchronize lip motion with audio, but only modify the mouth region while keeping head pose and expressions fixed. More recently, several approaches [3, 5, 12] leverage powerful video diffusion models to synthesize temporally coherent and fluid talking videos, yet their heavy computational cost limits real-time deployment. Apart from them, several works [30, 32] adopt a two-stage generation strategy, first producing an intermediate representation—such as facial landmarks or implicit keypoints [10]—and then rendering the final output conditioned on an identity image. This design achieves a strong balance between efficiency and flexibility.

In addition to speaking-oriented methods, a number of approaches focus on listening-head generation, which aims to produce non-verbal reactions—such as head nods or smiling expressions—driven by the speech of other speakers. For example, L2L [18] employs a codebook-based model to generate listening motions, while ELP [25] incorporates emotional cues derived from speech content. DIM [26] can transition between talking and listening states but cannot generate both states simultaneously, which limits its applicability for interactive head generation.

2.2 Conversational Talking Face Generation

Unlike single-person scenarios, conversational talking face generation focuses on modeling interactions among participants. Differing from early methods [18, 25, 26] that rely on manual role switching between speakers and listeners, INFP [35] is the first approach to generate interactive talking videos from dual audio streams, synthesizing motions conditioned on both speech and feedback without explicit role switching. ARIG [11] further extends it through an auto-regressive model. While these methods enable long multi-turn dialogues, they still suffer from limited partner-feedback modeling and remain constrained to single-person animation. Some 3D facial animation approaches, such as DualTalk [21] also attempt to model dialogues with motion feedback, but their mesh-based generation restricts their applicability to 2D video synthesis.

For multi-person video animation, recent works increasingly employ video generation models to synthesize multi-character conversations. For example, MultiTalk [14] adopts Wan 2.1 [29] as its backbone with a human-binding strategy to produce multi-person animations, but it cannot scale beyond two participants. Emerging systems such as Veo3 [9] and Sora2 [19] achieve impressive audio-visual joint generation but are computationally costly, remain closed sourced, and require massive data, which cannot be adopted for high-throughput application like companion. In contrast, our method explicitly models interactions among multi-participants and achieves efficient real-time animation.

3 Method

Our proposed framework InterTalk aims to generate conversational talking face that achieve flexibility, naturalness, and efficiency simultaneously. As illustrated in Fig. 2, we first apply a Responsive Context Encoder (RCE) to captures interactive cues from the environment, including the speech and motion feedback of other participants during conversation. An Interactive Motion Generator (IMG) then leverages these interaction features together with the self audio to generate corresponding motions. Finally, the Rendering Pipeline synthesizes talking videos for each participant based on the generated motions and composites them into a cohesive multi-person conversational scene. Each component plays a vital role and their detailed designs are described in the following sections.

3.1 Responsive Context Encoder

The RCE integrates environmental information into a highly interactive representation, providing feature F_{RCE} for motion generation in the IMG. Its overall process can be expressed as:

$$F_{\text{RCE}} = \text{RCE}(\{A_i, M_i\}_{i=1}^N), \quad (1)$$

where A_i and M_i denote the audio and motion inputs of the i -th participant, and N is the total number of participants. For each participant i , this module encodes both audio and motion inputs into a unified latent space. Specifically, the audio stream of participant i is encoded using Wav2Vec 2.0 [1] to obtain a robust speech representation F_a^i . Meanwhile, the motion input is disentangled into several facial components—lip motion M_i^{lip} , eyes motion M_i^{eye} , and head pose motion M_i^{pose} —each encoded separately and then combined into a comprehensive motion representation F_m^i to better capture non-verbal feedback within conversations. The encoding process is defined as:

$$F_a^i = \text{Enc}_a(A_i), \quad F_m^i = \text{Enc}_m(M_i^{\text{lip}}, M_i^{\text{eye}}, M_i^{\text{pose}}). \quad (2)$$

After obtaining the audio and motion features of each participant, we apply a cross-attention mechanism to align and fuse them across modalities, bridging the gap between the two feature domains. To further enhance temporal coherence, a bi-directional LSTM [24] is subsequently applied to capture both past and future conversational dynamics to get i -th participants interactive context feature F_{RCE}^i . The overall process for the i -th participant is formulated as:

$$F_{\text{RCE}}^i = \text{biLSTM}(\text{CrossAttn}(F_a^i, F_m^i)). \quad (3)$$

Note that the context window covers only about 3 seconds rather than a significant future buffer, which is suitable for practical deployment. Finally, the RCE aggregates interactive features across all participants via adaptive attention pooling. Specifically, the reference feature $\bar{F} = \frac{1}{N} \sum_{i=1}^N F_{\text{RCE}}^i$ is obtained by averaging all participant features, while attention weights are computed based on

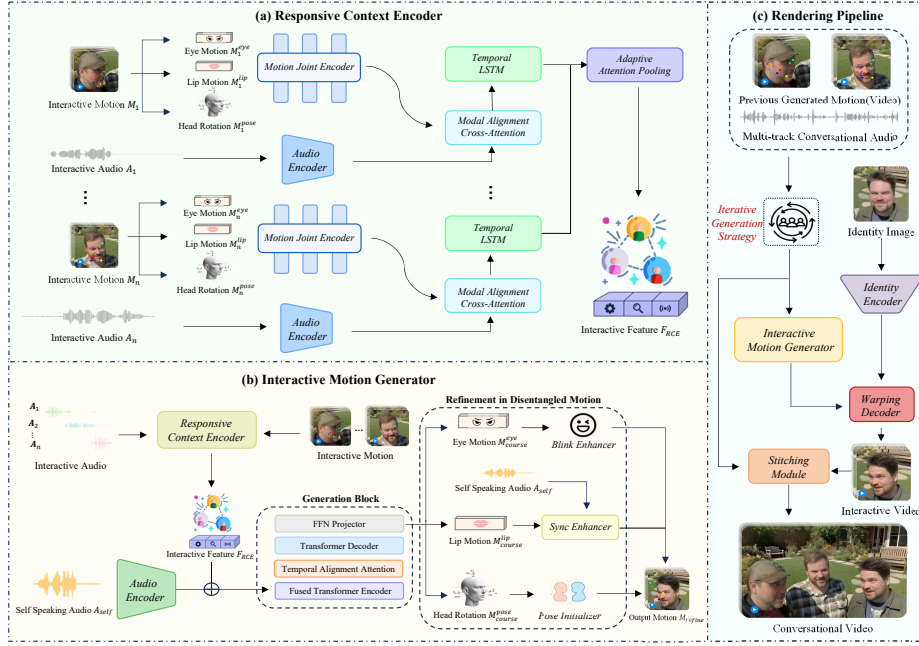


Fig. 2: Framework of InterTalk. InterTalk consists of three key components: a Responsive Context Encoder (RCE) that integrates environmental elements into an interactive feature representation; an Interactive Motion Generator (IMG) that produces fluid conversational motions; and a Rendering Pipeline that animates each participant to synthesise the final talking-face videos.

each participant’s similarity to this reference and dimension d of F_{RCE}^i , yielding higher weights for features closer to the conversational context. The process can be formulated as:

$$\alpha_i = \text{Softmax}\left(\frac{F_{RCE}^i \cdot \bar{F}}{\sqrt{d}}\right), F_{RCE} = \sum_{i=1}^N \alpha_i F_{RCE}^i. \quad (4)$$

In this way, the RCE produces a unified interactive context feature F_{RCE} that effectively captures cross-participant dependencies and provides global conversational context for downstream motion generation.

3.2 Interactive Motion Generator

After obtaining interactive features F_{RCE} from the RCE, the IMG generates temporally coherent and context-aware facial motions conditioned on both the self speaker’s audio and environmental interaction cues. The overall process can be formulated as:

$$M_{\text{refine}} = \text{IMG}(A_{\text{self}}, F_{RCE}), \quad (5)$$

where A_{self} denotes the self speaking audio, which is encoded into F_a using the same audio encoder as in the RCE to ensure feature-space consistency. The F_a is combined with the interactive feature F_{RCE} from Eqn. 1 before being fed into the motion generation network to produce M_{coarse} . It can be formulated as:

$$M_{\text{coarse}} = \text{GenBlock}(F_{\text{RCE}} + F_a). \quad (6)$$

The generation network comprises four sequential blocks. A fused Transformer Encoder [28] first integrates audio and interactive features within a unified latent space. Next, a temporal alignment attention module, adapted from FaceFormer [7], applies a causal mask to ensure temporal consistency and prevent future-frame leakage. A Transformer decoder then captures long-range dependencies and contextual transitions, followed by a feed-forward layer that projects the learned representations into subspaces of different facial components—lip, eye, and head pose—for targeted motion refinement.

Refinement on Disentangled Motion. We apply three modules to refine the decomposed motion from the coarse motion M_{coarse} . To improve lip-sync accuracy beyond the limitations of conversational data, we train a lightweight sync-enhancer on a large-scale single-speaker dataset. This module takes speech audio A_{self} and initial lip motion $M_{\text{coarse}}^{\text{lip}}$ as input and predicts local deformations for refined synchronization. It is trained independently after other modules are pretrained, ensuring compatibility with the full system. We further incorporate a Blink-Enhancer that regulates blink frequency f and noise z to produce natural eyelid closures. In addition, a Pose-Initializer adjusts the initial head orientation according to the relative facial position $(x_{\text{self}}, y_{\text{self}})$, maintaining mutual gaze alignment at the start of the interaction. The refinement process can be formulated as:

$$\begin{cases} M_{\text{refine}}^{\text{eye}} = \text{BlinkEnhancer}(M_{\text{coarse}}^{\text{eye}}, f, z), \\ M_{\text{refine}}^{\text{lip}} = \text{SyncEnhancer}(M_{\text{coarse}}^{\text{lip}}, A_{\text{self}}), \\ M_{\text{refine}}^{\text{pose}} = \text{PoseInitializer}(M_{\text{coarse}}^{\text{pose}}, x_{\text{self}}, y_{\text{self}}). \end{cases} \quad (7)$$

We then combine these refined components to obtain the final output M_{refine} . Further details on the target-specific enhancement are provided in the supplementary material.

3.3 Rendering Pipeline

While most prior approaches focus on generating a single speaker [11, 30, 32, 35], our rendering pipeline synthesizes coherent conversational videos with multiple participants. We adopt 3D implicit keypoints as intermediate representations, extracted through three modules: a pose estimator computing rotation R , translation t , and scale s ; an expression estimator predicting deformation δ ; and a canonical keypoint detector extracting canonical keypoints K_c from reference images I_{ref} . The original keypoints K_{ori} are obtained as:

$$K_{\text{ori}} = s \cdot (K_c \cdot R + \delta) + t. \quad (8)$$

The refined motion M_{refine} is decomposed into rotation R_d and expression deformation δ_d , which are then applied to transform K_{ori} into the driven implicit keypoints K_d . The process can be formulated as:

$$M_{\text{refine}} \Rightarrow \{R_d, \delta_d\}, K_d = s \cdot (K_c \cdot R_d + \delta_d) + t. \quad (9)$$

For each participant, the region is cropped from I_{ref} and encoded into an appearance embedding f_a via an identity encoder. A warping decoder fuses f_a with the motion sequence K_d predicted by the IMG to estimate a dense warping flow field. This flow is then applied to f_a to generate the final rendered frame with accurate alignment and natural motion:

$$I_{res} = \text{Decoder}(\text{Warp}(f_a, K_{ori}, K_d)). \quad (10)$$

Finally, we introduce a stitching module that spatially aligns and blends individual renderings. It refines local misalignment via fine-grained keypoint deformations and uses a spatial blending mask for smooth transitions. This allows InterTalk to seamlessly composite all animated participants into the original scene, producing visually coherent multi-participant animations. Further details are provided in the supplementary material.

3.4 Training and Inference

Loss Function. We use the L2 loss $\mathcal{L}_{\text{rec}}$ and velocity loss $\mathcal{L}_{\text{vel}}$ as supervision on motion, which can be formulated as:

$$\mathcal{L}_{\text{motion}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{vel}} \mathcal{L}_{\text{vel}}. \quad (11)$$

This loss ensures motion fidelity and temporal smoothness across frames. In addition, the sync-enhancer is trained on single-person talking-face data with a synchronization loss $\mathcal{L}_{\text{sync}}$ [22] to refine the lip-audio alignment between speech and facial motion. Further training details and explanations are provided in the supplementary material.

Iterative Generation Strategy. During inference, the model can only utilize motion feedback from previously generated participants rather than all individuals, which may compromise interactive dynamics. To mitigate this limitation, we introduce an iterative motion generation strategy that progressively incorporates feedback across cycles. As illustrated in Fig. 3, we sequentially generate motions by using the outputs from the previous iteration as inputs for the current one, continuously updating a shared motion buffer. Through several iterations, each participant’s motion gradually incorporates feedback from others, resulting in more responsive and natural interactions without delay. Formally, let M_{ij} denote the motion of the i -th participant at iteration j . The iterative process can be defined as:

$$M_{ij} = \text{IMG}\left(A_i, \text{RCE}\left(M_{(j-1)}^{-i}, A_{(j-1)}^{-i}\right)\right), \quad (12)$$

where A_i represents the audio feature of the i -th participant, and RCE encodes the interaction context from the previous iteration’s motions $M_{(j-1)}^{-i}$ and audio features $A_{(j-1)}^{-i}$ of all other participants except themselves ($\neg i$). The initial

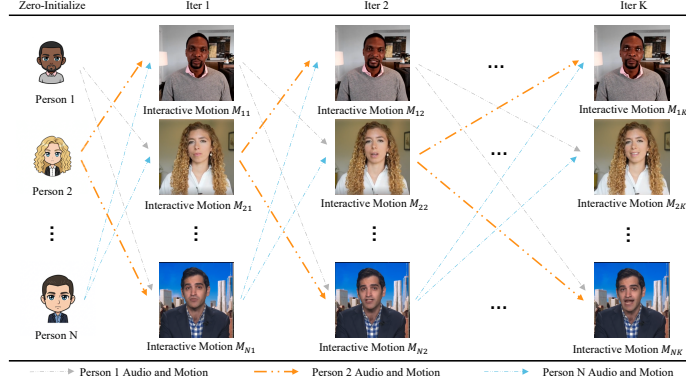


Fig. 3: Iterative Generation Strategy. By progressively updating motions through multiple iterations, the final results accurately capture mutual feedback among all participants.

feedback motions M_{i0} and audio features A_0^i are zero-initialized at the start of inference. During the iterative process, the three target-refinement modules in IMG are disabled. Since the IMG is lightweight and computationally efficient, this strategy introduces negligible overhead compared with the overall runtime, preserving real-time generation capability.

Video Dubbing and Single-role Generation. Our method can be extended to video inputs by processing each frame as K_{ori} , enabling video dubbing with interactive conversations while preserving natural body motion and background changes. For single-role interactive head generation, we disable pose initialization and iterative generation strategy, directly animating from multi-track audio and motion feedback when available.

4 Data Construction

Existing public datasets for interactive talking faces remain limited in quality and conversational depth. For instance, ViCo [34] focuses on single-round dialog, while DualTalk [21] provides only 3D FLAME [16] coefficients. Moreover, most multi-person datasets concatenate separate single-speaker recordings instead of capturing participants in a shared scene. To overcome these issues, we build a new multi-person conversational dataset and enrich it with 3D data by converting FLAME coefficients into motions for our framework.

2D Data Collection. As shown in Fig. 4(a), we collect multi-speaker conversation videos from the Internet and manually segment them to ensure that each clip includes at least two participants continuously appearing in the same scene. After cleaning, the dataset contains 143 identities and 205 clips. To separate individual audio tracks from mixed recordings, we employ a visual-audio separation network [15] that leverages lip movements to infer each speaker’s voice as illustrated in Fig. 4(b). Additional TTS-based audio segments are synthesized to enhance diversity during evaluation.

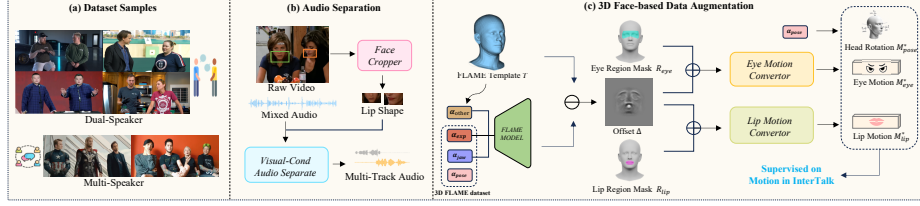


Fig. 4: Dataset Information. (a) Samples from InterTalk dataset. (b) The processing pipeline for collected videos and multi-track audio separation. (c) 3D face-based data augmentation strategy using 3D conversation datasets, generating motions from FLAME coefficients.

3D Face-based Data Augmentation. To further enhance conversational diversity and interactive learning, we propose a 3D face-based augmentation strategy that integrates DualTalk’s [21] 3D face data into training. As shown in Fig. 4(c), the 3D FLAME [16] coefficients are converted into motion signals through two converter networks trained on a 2D talking-face dataset [33], where the FLAME parameters are annotated using EMOCA [6]. Specifically, we reconstruct a 3D mesh using the jaw pose α_{jaw} and expression α_{exp} , while keeping the other FLAME coefficients α_{other} identical to the template mesh T . The vertex offsets Δ are then obtained by subtraction, formulated as:

$$\Delta = \text{FLAME}(\alpha_{\text{exp}}, \alpha_{\text{jaw}}, \alpha_{\text{other}}) - T. \quad (13)$$

Afterward, we apply eye and lip region masks R_{eye} and R_{lip} to extract localized expressions, which are fed into corresponding converters to generate motion features as:

$$M_c^* = \text{Convertor}_c(R_c(\Delta)), \quad c \in \{\text{eye}, \text{lip}\}. \quad (14)$$

Head-pose information is directly mapped to the motion M_{pose}^* . This converted dataset further enhances the realism and fluidity of conversational behavior, improving both inter-participant responsiveness and motion coherence.

5 Experiments

5.1 Experimental Settings

Implementation Details. We preprocess the dataset by converting videos to 25 fps and resampling audio at 16kHz. The RCE and IMG modules are trained

Table 1: Quantitative results with state-of-the-art talking face generation methods on HDTF [33] dataset.

Method	Sync-C $\uparrow$	Sync-D $\downarrow$	FID $\downarrow$	NIQE $\downarrow$	FVD $\downarrow$
SadTalker [32]	7.15	7.93	40.75	46.30	291.66
Hallo2 [5]	7.53	7.97	33.15	14.25	205.60
EchoMimic [3]	5.94	9.11	<u>28.13</u>	<u>13.42</u>	284.38
Sonic [12]	<u>8.26</u>	6.88	34.41	13.88	<u>204.55</u>
Ours	8.30	<u>6.91</u>	23.07	13.29	178.73

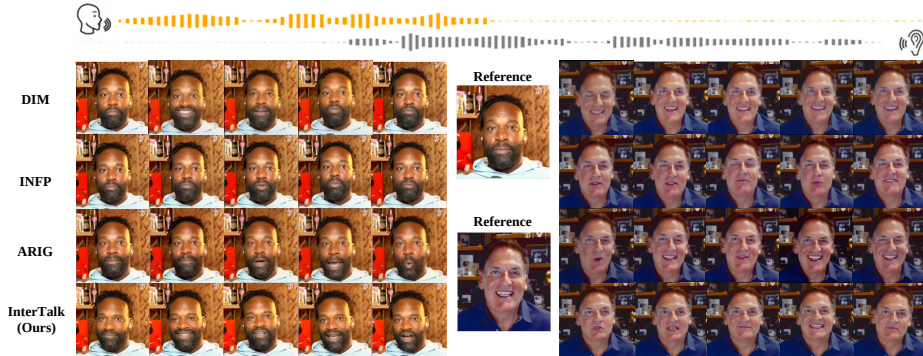


Fig. 5: Comparison of Interactive Head Generation. Two identities are sampled from the DyConv [35] dataset proposed by INFP, which is not publicly available and therefore lacks ground-truth results.

using the Adam optimizer [13] with a learning rate of 1×10^{-4} . The components of the MAP are initialized from LivePortrait [10] to leverage its robust rendering quality. Training is performed on eight RTX 4090 GPUs for three days. During inference, each participant is animated in parallel across multiple GPUs to achieve faster generation speed. Our method produces 512×512 resolution outputs before stitching them back into the original frame. Additional implementation details are provided in the supplementary material.

Metrics. We evaluate our method across various scenarios, covering both speaking and interactive aspects. For speaking evaluation, we assess lip synchrony using Sync-C and Sync-D from SyncNet [4], image quality with FID [23] and NIQE [17] for reference and no-reference assessment, and temporal consistency with FVD [27]. For interactive evaluation, we extract 3D FLAME [16] coefficients from generated videos and perform quantitative analysis using multiple metrics, including Fréchet Distance (FD), Paired Fréchet Distance (P-FD), Mean Squared Error (MSE), Similarity Index for Diversity (SID), and Residual Pearson Correlation Coefficient (rPCC). We further report generation speed in frames per second (fps) to measure system efficiency.

5.2 Comparison

Talking Face Generation. We first evaluate our method on general single person talking face generation using metrics that assess image quality and lip synchronization. To adapt our framework on this task, we set the interactive audio and motion feedback inputs to zero. We compare our approach with four advanced baselines, including SadTalker [32], EchoMimic [3], Hallo2 [5], and Sonic [12] on the HDTF [33] dataset. As shown in Tab. 1, our method demonstrates superior performance in both video quality and lip synchronization. We achieve better FID and NIQE scores owing to the robust rendering pipeline, and obtain the best SyncNet score, slightly outperforming Sonic thanks to the

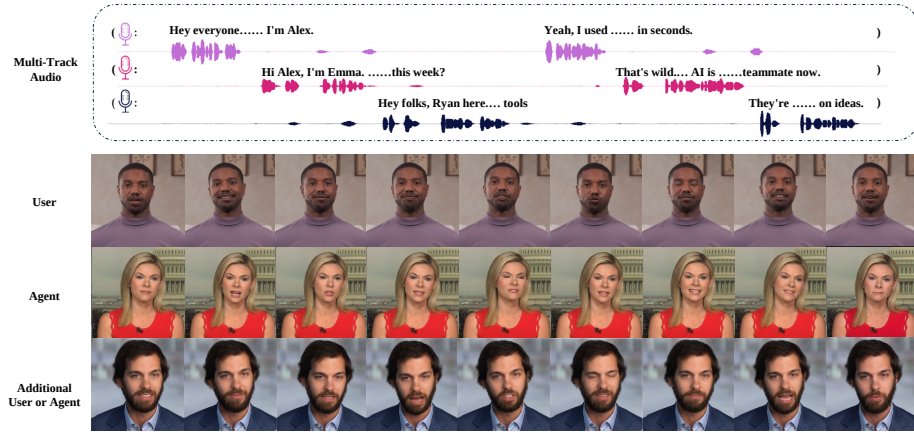


Fig. 6: Interactive Head Generation. Our method generates highly responsive interactive agent avatars and supports additional user motion inputs or agent outputs from multi-track audio, as InterTalk handles an arbitrary number of participants.

integration of the sync-enhancer within the IMG. The higher FVD score further indicates that our method achieves better temporal consistency. The qualitative results are demonstrated in supplementary.

Interactive Head Generation. We evaluate our method against existing other interactive head generation approaches, including DIM [26], INFP [35], and ARIG [11]. For quantitative evaluation, we conduct experiments on the ViCo [34] dataset for listening-head generation and adopt the metrics reported in the original papers due to the lack of official implementations. As shown in Tab. 2, our method achieves superior accuracy and diversity, demonstrating its clear advantage in interactive head generation. In terms of visual results, since INFP and ARIG have not released their source code, we compare our results using demo videos available on their project pages. As shown in Fig. 5, our approach produces natural and expressive head movements with sharper facial details and better identity preservation. In contrast, DIM and INFP display relatively rigid expressions, while ARIG maintains an almost static head pose. In comparison,

Table 2: Quantitative results with state-of-the-art interactive head generation methods on ViCo [34] dataset.

Methods	FD ↓		RPCC ↓		SID↑		MSE↓	
	exp	pose	exp	pose	exp	pose	exp	pose
DIM [26]	23.88	0.06	0.06	0.03	3.71	2.35	0.70	0.02
INFP [35]	18.63	0.07	-	-	4.78	3.92	<u>0.51</u>	0.01
ARIG [11]	<u>18.39</u>	0.05	0.05	0.01	<u>4.82</u>	<u>3.94</u>	-	-
DualTalk [21]	22.27	0.04	0.07	0.03	4.36	3.84	0.58	0.01
Ours	18.33	0.04	0.05	0.01	4.85	4.07	0.48	0.01

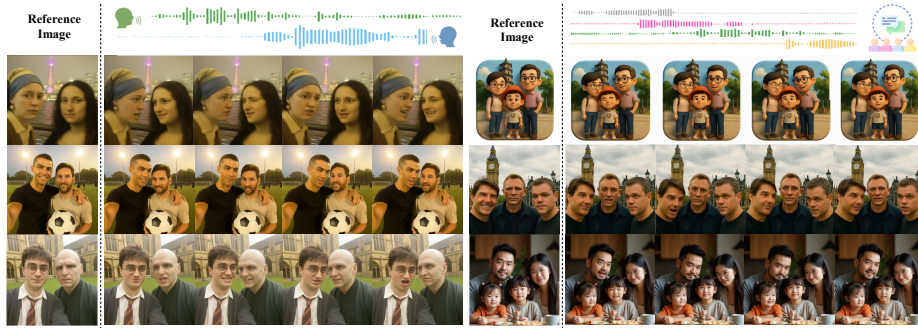


Fig. 7: InterTalk for Multi-person Conversation. Our method supports diverse input styles and imposes no constraints on the number of participants, while producing highly interactive and realistic conversational behaviors.

our method generates more context-aware gestures—such as gentle nods and natural smiles—demonstrating strong responsiveness during interaction. Furthermore, as shown in Fig. 6, our method generates lifelike virtual agents that respond naturally to users. Since it supports an arbitrary number of participants, both the number of output agents and input users are unrestricted, enabling applications like multi-participants remote meeting.

Conversations with Multi-Participants. Due to the scarcity of publicly available multi-person conversation video and related methods, we evaluate our approach on our collected dataset with MultiTalk [14], as it is the only open source method for multi-person talking face generation. Our method outperforms MultiTalk on nearly all metrics except SID, likely due to the stronger generalization ability of large video-generation models. Moreover, our approach achieves real-time FPS, while MultiTalk requires lengthy inference.

As shown in Fig. 7, our method not only generates natural and engaging interactions between participants but also produces multi-participant conversational videos without any restricted on the number of participants. Additional visual comparisons with MultiTalk are provided in the supplementary material. In summary, our method demonstrates strong interactive capability and high efficiency in conversational talking-face generation.

Table 3: Quantitative comparisons and ablation study on multi-person conversation generation on InterTalk Dataset.

Methods	Sync-C $\uparrow$	FD $\downarrow$	RPCC $\downarrow$	SID $\uparrow$	FPS $\uparrow$
MultiTalk [14]	7.94	41.74	0.24	5.14	0.61
Ours	8.30	22.67	0.09	4.28	31.23
<i>w/o</i> Motion Feedback in RCE	8.13	29.72	0.15	2.71	31.85
<i>w/o</i> Iterative Generation Strategy	8.25	24.26	0.11	4.12	34.67
<i>w/o</i> Disentanglement on Motion	7.58	25.17	0.11	3.85	31.56
<i>w/o</i> Data Augmentation	7.01	38.67	0.20	1.95	31.23

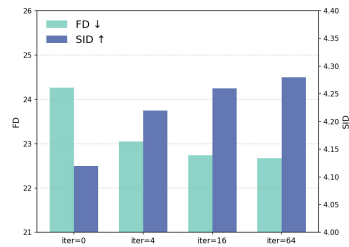


Fig. 8: Variation of FD and SID across iterations.

Table 4: User study on Interactive Head Generation. Best results are highlighted in bold.

Methods	Lip-sync Accuracy	Interaction Fluency	Motion Naturalness	Visual Realism
DIM [26]	1.13	1.24	1.19	1.07
INFP [35]	2.93	3.85	3.21	3.59
ARIG [11]	3.87	2.59	2.13	2.48
Ours (Single)	4.03	3.96	4.55	4.12

5.3 Ablation Study

As shown in Tab. 3, we conduct an ablation study on four key components: the motion feedback in RCE, the iterative generation strategy, the disentanglement on motion refinement in IMG, and the 3D face-based data augmentation. Removing motion feedback in RCE leads to a noticeable drop in interaction quality, as reflected by the interaction matrix on the FLAME coefficients. Introducing the iterative generation strategy significantly improves interaction performance with only a minor decrease in speed, as the model progressively refines motions to reduce response delays instead of relying solely on previously generated frames. The disentanglement allows the sync-enhancer to further boost Sync-C scores for lip synchronization by leveraging a large-scale single-speaker talking-face dataset to compensate for the limited diversity of conversational data. Moreover, the 3D face-based data augmentation greatly enhances interactive performance. Overall, the results show that each component contributes effectively to the framework.

As shown in Fig. 8, we analyze how FD and SID vary with different iteration numbers in the iterative generation strategy. The results demonstrate that this strategy progressively improves both motion realism and diversity. Specifically, the FD steadily decreases with more iterations, indicating that the generated motion distribution becomes increasingly similar to real motion data. At the same time, the SID rises across iterations, demonstrating that the iterative refinement enhances motion variability and expressiveness. These results confirm the effectiveness of our iterative generation strategy in producing natural and diverse conversational behaviors.

5.4 User Study

To evaluate perceptual quality, we conducted a user study on interactive head generation following the MOE protocol with 23 participants. As shown in Tab. 4, each method was rated on four aspects: lip-sync accuracy, interaction fluency, motion naturalness, and visual realism—using a 1–5 Likert scale where higher scores indicate better quality. Our method outperforms all baselines across every metric, demonstrating clear advantages in both realism and interactivity. Participants consistently preferred our results, praising the smoothness of facial motion and the natural transitions in interactive responses. In particular, our approach achieved the highest scores in motion naturalness and interaction fluency,

verifying the effectiveness of our iterative generation strategy. Furthermore, the superior lip-sync accuracy highlights the strength of our sync enhancer module in producing temporally aligned mouth movements. These results collectively confirm that InterTalk delivers the most lifelike and engaging interactive head generation among all compared methods.

6 Conclusion

In this paper, we present InterTalk, a novel and efficient framework for highly interactive conversational talking-face generation without constraints on the number of participants. The framework employs a Responsive Context Encoder (RCE) to encode audio and motion feedback into interactive features, ensuring flexibility. An Interactive Motion Generator (IMG) then synthesizes conversational motions with disentangled refinement, while a Rendering Pipeline animates multiple participants in real-time. We further incorporate an iterative generation strategy that refines motion feedback across participants, promoting coherent and responsive interactions. To address the scarcity of multi-person conversational data, we construct a new dataset and introduce a 3D face-based data augmentation strategy to enhance training. Extensive experiments demonstrate that InterTalk achieves flexible, natural, and efficient performance in interactive conversational talking face generation.

Acknowledgments

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0123501), Beijing Natural Science Foundation L242092, Chinese National Natural Science Foundation Projects 92570119, 62276254, U23B2054, the Science and Technology Development Fund of Macau Project 0140/2024/AGJ, and InnoHK program.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* **33**, 12449–12460 (2020)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023)
3. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 2403–2410 (2025)
4. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: *Workshop on Multi-view Lip-reading, ACCV* (2016)

5. Cui, J., Li, H., Yao, Y., Zhu, H., Shang, H., Cheng, K., Zhou, H., Zhu, S., Wang, J.: Hallo2: Long-duration and high-resolution audio-driven portrait image animation. arXiv preprint arXiv:2410.07718 (2024)
6. Daněček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022)
7. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18770–18780 (2022)
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014)
9. Google: Veo 3. <https://aistudio.google.com/models/veo-3> (2025), <https://aistudio.google.com/models/veo-3>
10. Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., Zhang, D.: Liveportrait: Efficient portrait animation with stitching and retargeting control. arXiv preprint arXiv:2407.03168 (2024)
11. Guo, Y., Liu, X., Zhen, C., Yan, P., Wei, X.: Arig: Autoregressive interactive head generation for real-time conversations. arXiv preprint arXiv:2507.00472 (2025)
12. Ji, X., Hu, X., Xu, Z., Zhu, J., Lin, C., He, Q., Zhang, J., Luo, D., Chen, Y., Lin, Q., et al.: Sonic: Shifting focus to global audio perception in portrait animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 193–203 (2025)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kong, Z., Gao, F., Zhang, Y., Kang, Z., Wei, X., Cai, X., Chen, G., Luo, W.: Let them talk: Audio-driven multi-person conversational video generation. arXiv preprint arXiv:2505.22647 (2025)
15. Li, K., Yang, R., Sun, F., Hu, X.: Iianet: an intra-and inter-modality attention network for audio-visual speech separation. In: Proceedings of the 41st International Conference on Machine Learning. pp. 29181–29200 (2024)
16. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **36**(6), 194:1–194:17 (2017), <https://doi.org/10.1145/3130800.3130813>
17. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012)
18. Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., Ginosar, S.: Learning to listen: Modeling non-deterministic dyadic facial motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20395–20405 (2022)
19. OpenAI: Sora: Creating video from text. <https://sora.chatgpt.com> (2024), <https://sora.chatgpt.com>, accessed: 2024-11-05
20. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
21. Peng, Z., Fan, Y., Wu, H., Wang, X., Liu, H., He, J., Fan, Z.: Dualtalk: Dual-speaker interaction for 3d talking head conversations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21055–21064 (2025)
22. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia. pp. 484–492 (2020)

23. Seitzer, M.: pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid> (August 2020), version 0.3.0
24. Sherstinsky, A.: Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena* **404**, 132306 (2020)
25. Song, L., Yin, G., Jin, Z., Dong, X., Xu, C.: Emotional listener portrait: Neural listener head generation with emotion. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 20839–20849 (2023)
26. Tran, M., Chang, D., Siniukov, M., Soleymani, M.: Dim: Dyadic interaction modeling for social behavior generation. In: *European Conference on Computer Vision*. pp. 484–503. Springer (2024)
27. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
29. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025)
30. Wang, B., Zhu, X., Shen, F., Xu, H., Lei, Z.: Pc-talk: Precise facial animation control for audio-driven talking face generation. *arXiv preprint arXiv:2503.14295* (2025)
31. Wang, D., Dai, B., Deng, Y., Wang, B.: Disentangling planning, driving and rendering for photorealistic avatar agents. In: *European Conference on Computer Vision*. pp. 137–147. Springer (2024)
32. Zhang, W., Cun, X., Wang, X., Zhang, Y., Shen, X., Guo, Y., Shan, Y., Wang, F.: Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8652–8661 (2023)
33. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3661–3670 (2021)
34. Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T., Mei, T.: Responsive listening head generation: a benchmark dataset and baseline. In: *European conference on computer vision*. pp. 124–142. Springer (2022)
35. Zhu, Y., Zhang, L., Rong, Z., Hu, T., Liang, S., Ge, Z.: Infp: Audio-driven interactive head generation in dyadic conversations. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 10667–10677 (2025)